# Giovanni in the Cloud: Earth Science Data Exploration in Amazon Web Services

IN23A-0080

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

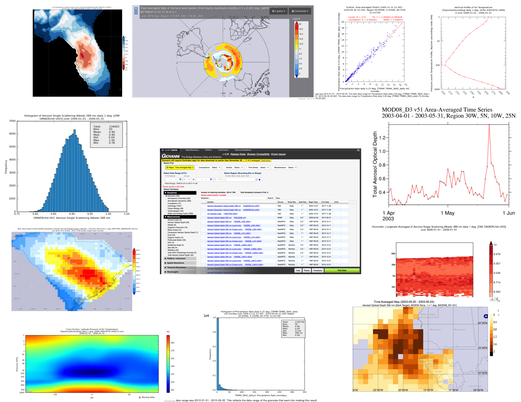
**Maksym Petrenko[1,2], Mahabal Hegde[1,2], Christine Smit[1,3], Hailiang Zhang[1,2], Paul Pilone[4], Andrey Zasorin[1,3] , Long Pham[1]**

[1] NASA Goddard Space Flight Center, [2] ADNET Systems Inc., [3] Telophase Corp, [4] Element84

## What is Giovanni?

Giovanni (https://giovanni.gsfc.nasa.gov) is an online tool for exploration of geo-spatial data with:

- Twenty-two (22) analysis and visualization services at the click of a button
- Access to over 1600 data variables
- Persistent URLs for sharing data and visualizations




## Notable Features

- Single page application for specifying service parameters, navigating and manipulating results
- Rapid exploration of geo-spatial data in time and space
- Serves broad spectrum of users from students to subject matter experts

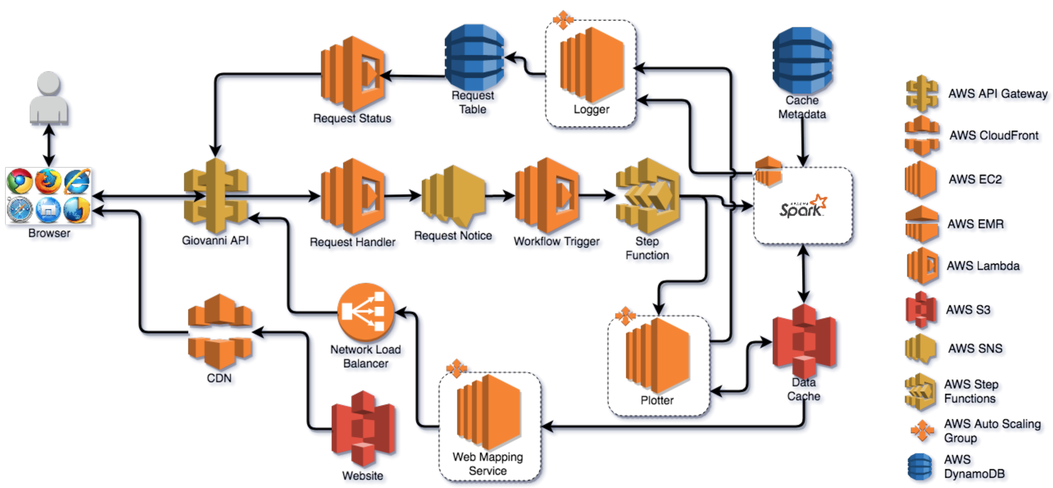
## Pain Points

- Emphasis on feature set rather than reliability and performance, the two key pillars of a well architected framework
- Victim of its own success; unable to meet spikes in demand during training, and “seasonal” events such as conferences and end of academic terms
- Increased demand on resources due to higher resolution data and user demand for data statistics

## Leveraging Cloud

- Server-less architecture: AWS-managed solutions for services where possible.
  - AWS API Gateway for service endpoints
  - AWS Lambda, Simple Queueing Service (SQS), Simple Notification Service (SNS) for triggering request processing
  - AWS Simple Storage Service (S3) for webhosting and data storage
  - AWS Elastic MapReduce (EMR) for cluster computing
  - AWS Elastic Compute Cloud (EC2) for general computing (Example: Web Mapping Service)
- Micro services: each operation is an independent service, making chaining of services feasible
- OpenAPI based service specifications: enables language agnostic service definition
- Auto-scaling: to meet demand spikes and compute-intensive services
- Use of Apache Parquet, a columnar and open-source data store widely supported in the Big Data community, for storing analysis-ready data
- Built-in synergistic evolution: as AWS evolves in hardware and software, Giovanni benefits

## Architecture




## See also

IN31A-0068: Use of Schema on Read in Earth Science Data Archives
IN41B-0039: The Value of Data and Metadata Standardization for Interoperability in Giovanni

## Prototyped Features

- Services
  - Time-averaged map: most popular service in Giovanni
  - Area-averaged time series: most resource intensive service in Giovanni
- Plot-centric instead of variable-centric user interface: users add data to plots simplifying user experience

## Performance Analysis

| | On-premises Version | Cloud Version (estimated) |
|---|---|---|
| Data Cache Size (Approximately 1600 variables) | 30 TB | 24.8 TB |

Compute Node: Single m4.10xlarge EC2

| Service | Data | Time HH:MM:SS (On-premises) | Time HH:MM:SS (Cloud) |
|---|---|---|---|
| Area-Averaged Time Series | 37+ years of Global $0.5^0$ x $0.67^0$ hourly-data | Not Available | 00:22:58 |
| Time-Averaged Map | 14 years of 20 x 20 grid $0.25^0$ daily data | 00:04:52 | 00:02:52 |
| Time-Averaged Map | 14 years of Global $0.25^0$ daily data | 00:09:03 | 00:04:00 |

## Cost Analysis

| | | |
|---|---|---|
| Storage | 25TB | $600 per month |
| Compute cost | 2500 requests/day | $1600 per month |